

# Open (Clinical) LLMs are Sensitive to Instruction Phrasings

Alberto Mario Ceballos Arroyo\*<sup>γ</sup> Monica Munnangi\*<sup>γ</sup> Jiuding Sun<sup>γ</sup>  
Karen Y.C. Zhang<sup>γ</sup> Denis Jered McInerney<sup>γ</sup>◇ Byron C. Wallace<sup>γ</sup> Silvio Amir<sup>γ</sup>

<sup>γ</sup>Northeastern University ◇Codametrix

{ceballosarroyo.a, munnangi.m, sun.jiu, zhang.yuchen, b.wallace,s.amir}@northeastern.edu

jmcinerney@codametrix.com

## Abstract

Instruction-tuned Large Language Models (LLMs) can perform a wide range of tasks given natural language instructions to do so, but they are sensitive to how such instructions are phrased. This issue is especially concerning in healthcare, as clinicians are unlikely to be experienced prompt engineers and the potential consequences of inaccurate outputs are heightened in this domain.

This raises a practical question: *How robust are instruction-tuned LLMs to natural variations in the instructions provided for clinical NLP tasks?* We collect prompts from medical doctors across a range of tasks and quantify the sensitivity of seven LLMs—some general, others specialized—to natural (i.e., non-adversarial) instruction phrasings. We find that performance varies substantially across all models, and that—perhaps surprisingly—domain-specific models explicitly trained on clinical data are especially brittle, compared to their general domain counterparts. Further, arbitrary phrasing differences can affect fairness, e.g., valid but distinct instructions for mortality prediction yield a range both in overall performance, and in terms of differences between demographic groups.

## 1 Introduction

Modern LLMs—e.g. GPT-3.5+ (Radford et al., 2019; Ouyang et al., 2022), the FLAN series (Chung et al., 2022), Alpaca (Taori et al., 2023), Mistral (Jiang et al., 2023)—can execute arbitrary tasks *zero-shot*, i.e., provided with only instructions rather than explicit training examples. LLMs have also shown promising improvements in performance on classification and information extraction (IE) tasks, such as named entity recognition (Brown et al., 2020; Munnangi et al., 2024) and relation extraction (Wadhwa et al., 2023a; Ashok and Lipton, 2023; Jiang et al., 2024) in both general and specialized domains like biomedical and

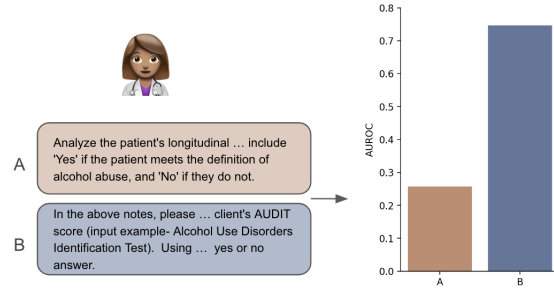


Figure 1: How much does LLM performance on clinical tasks depend on the arbitrary phrasings of instructions? Here we show an illustrative example: Discrepancy in AUROC score for CLINICAL CAMEL on the cohort selection-alcohol abuse classification task, when given the worst (A) and the best (B) performing prompts for ALCOHOL-ABUSE classification task.

scientific literature (Agrawal et al., 2022; Wadhwa et al., 2023b; Asada and Fukuda, 2024).

However, prior work has shown that LLMs do not “understand” prompts (Webson and Pavlick, 2022) and are sensitive to the particular phrasings of instructions (Lu et al., 2022; Sun et al., 2023). Domain experts in specialized domains such as medicine are especially likely to interact with models by providing instructions (i.e., in *zero-shot* settings), and are unlikely to be talented prompt engineers. For instance, a clinician might task a model to “Extract and summarize the findings of the patient’s last X-ray”, or ask “When did the patient last receive a painkiller?”. It is unrealistic to fine-tune models for every possible such task; hence the appeal of models responsive to arbitrary prompts. A downside, however, is that a clinician’s particular phrasing may dramatically affect model performance (Figure 1). Such unpredictability is especially troublesome in healthcare, where poor performance might ultimately impact patient health.

In this work we ask: **How sensitive are LLMs—general and domain-specific—to plausible instruction phrasing variations for clinical tasks?**

\*Equal contribution

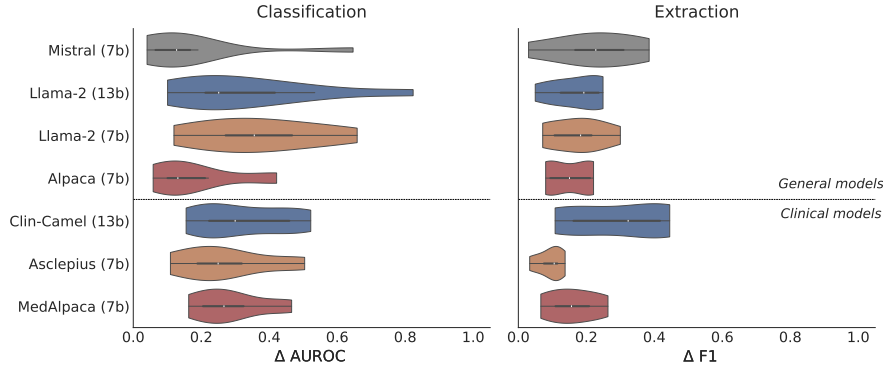


Figure 2: Variance in performance for clinical classification and information extraction tasks for each model. We show the distribution of **deltas between the best and worst performing prompt** for each task.

Our analysis deepens prior work on robustness by focusing on the clinical domain; this is important both due to the higher stakes and because clinical notes differ qualitatively from general domain text. For example, notes in EHR often contain grammatical errors (“*Pt complains of headache, and feel dizzy.*”); abbreviations not defined in context (“*Pt*” could be “*patient*” or “*Prothrombin time*”), and; domain-specific jargon (“*edema*”, “*Diuretic*”).

Therefore, one of the key aspects we consider is the domain-specificity of models. Are clinical LLMs more (or less) robust to different valid instruction phrasings written by doctors, compared to their general domain counterparts? To assess this, we evaluate recently released LLM variants trained on synthetic datasets comprising automatically generated clinical notes (Kweon et al., 2023), and medical dialogue from case reports found in biomedical literature (Toma et al., 2023). We find that performance varies substantially given alternative instruction phrasings for both general and clinical LLMs. Figure 2 shows the distribution of deltas between the best and worst performing prompts across a set of clinical classification and information extraction tasks.

Finally, we investigate how instruction phrasings impact the fairness of predictions, by which here we mean observed differences in performance between demographic subgroups. The degree to which LLMs might perpetuate and exaggerate such disparities in clinical use is a topic of active research (Omiye et al., 2023; Pal et al., 2023; Zack et al., 2024). Here we contribute to this by investigating the interaction between prompt phrasings and fairness. We find significant performance differences (up to 0.35 absolute difference in AUROC) in a mortality prediction task from MIMIC-III between *White* and *Non-White* subgroups and also

a significant disparity between *Male* and *Female* patients (up to 0.19 absolute difference in AUROC). To facilitate future research in this direction, we release our code and prompts<sup>1</sup>.

## 2 Experimental Framework

Our experimental setup is intended to quantify the robustness of LLMs to natural variations in instructional phrasings for clinical tasks. We considered a set of ten clinical classification tasks and six information extraction tasks drawn from MIMIC-III (Johnson et al., 2016) and prior i2b2 and n2c2 challenges,<sup>2</sup> summarized in Table 1 (§2.1). We recruited a diverse group of medical professionals to write prompts for each task (§2.2). We then evaluated the performance, variance, and fairness of seven LLMs (four general-domain and three domain-specific) across prompts (§2.3).

### 2.1 Tasks and Datasets

MIMIC-III (Johnson et al., 2016) is a database of de-identified EHR comprising over 40k patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. It comprises structured variables and clinical notes (e.g., doctor and nursing notes, radiology reports, discharge summaries); we focus on the latter. MIMIC-III also contains demographic information, including ethnicity/race, sex, spoken language, religion, and insurance status (Chen et al., 2019). As an illustrative predictive task, we consider in-hospital mortality prediction, which has been the subject of prior work (Harutyunyan et al., 2017). Owing to compute constraints, we sub-sampled the test-split to 10% of the data (preserving class ratio), yielding 160 records for evaluation.

<sup>1</sup><https://github.com/alceballosa/clin-robust>

<sup>2</sup><https://n2c2.dbmi.hms.harvard.edu/>

Dataset	TASK	TEST SET	TASK TYPE
MIMIC-III	In-hospital Mortality	160	Binary Classification
Obesity co-morbidity	Asthma	507	Binary Classification
	CAD	507	Binary Classification
	Diabetes	507	Binary Classification
	Obesity	507	Binary Classification
Cohort Selection	Abdominal	86	Binary Classification
	Alcohol-Abuse	86	Binary Classification
	Drug-Abuse	86	Binary Classification
	English	86	Binary Classification
	Decisions	86	Binary Classification
Medical Challenge	Medication	251	Extraction
	Concept Problem	256	Extraction
Relation Challenge	Concept Test	256	Extraction
	Concept Treatment	256	Extraction
Adverse Drug Effects	Drug	202	Extraction
Risk Assessment	Risk Factor CAD	514	Extraction

Table 1: Tasks and datasets used for evaluation.

**n2c2 2018 Cohort Selection Challenge (Stubbs and Uzuner, 2019)** aims to identify whether a patient meets the criteria for inclusion in a clinical trial based on their longitudinal records. The dataset contains 288 patients, their associated clinical notes and a set of binary labels indicating whether they meet the criteria for each of 13 possible cohorts (e.g., drug abuse, alcohol abuse, ability to make decisions, among others). In this study, we focus on the 5 cohorts shown in Table 1 and treat each as an independent binary classification task aiming to predict whether the criteria is “met” or “not met”.

**i2b2 2008 Obesity Challenge (Uzuner, 2009)** entails identifying patients suffering from obesity and its co-morbidities from their discharge summary notes. The dataset comprises 1027 pairs of de-identified discharge summaries and 16 disease labels from intuitive judgements which are based on the entire discharge summary. We report the performance for obesity and three co-morbidities (i.e., asthma, atherosclerotic cardiovascular disease (CAD), and diabetes mellitus (DM)), each framed as a binary classification task aiming to predict whether the condition is “present” or “absent”.

**n2c2 2018 Adverse Drug Events and Medication Extraction in EHRs (Henry et al., 2020)** consists of a relation extraction task focused on identifying drugs/medications and their relations to

adverse events for the patient. The dataset contains 202 patients and we focus only on the named entity recognition portion of the task (i.e. recognizing spans referring to drugs/medications).

**i2b2 2014 Identifying Risk Factors for Heart Disease over Time (Stubbs et al., 2015):** entails identifying medical risk factors linked to Coronary Artery Disease (CAD) in the EHR of patients with diabetes. The target factors include hypertension, obesity, smoking status, diabetes, hyperlipidemia, family history, and CAD itself. Here we consider only the latter.

**i2b2 2010 Relations Challenge (Uzuner et al., 2011)** consists of three related tasks: (1) identification of medical problems, tests, and treatments; (2) classification of assertions made on medical problems; and (3) relation extraction concerning medical problems, tests, and treatments. The data for this challenge includes discharge summaries from Partners HealthCare, and the Beth Israel Deaconess Medical Center (Lee et al., 2011), as well as discharge summaries and progress notes from the University of Pittsburgh Medical Center. We conduct evaluation on the first task (i.e. extraction of problems, tests, and treatments) over the notes of 256 patients.

**i2b2 2009 Medication Extraction Challenge (Patrick and Li, 2010)** focuses on the extraction of medications from clinical notes in the EHR,

as well as their modes, reasons and frequency of administration. We center our analysis on medication extraction only, which encompasses around 1250 unique medications over 251 notes.

## 2.2 Instruction Collection

We hired twenty medical professionals from different professional and demographic backgrounds, with varying medical specialties and years of experience. These included medical doctors (physicians, surgeons), medical writers/editors, nurses, and medical consultants from various countries, such as the United States, Nigeria, Kenya, Canada, Zambia, Egypt, Malawi, Pakistan, Philippines, and Ethiopia. All participants were either native-speakers or proficient in English. It should also be noted that participants were not required to have experience with LLMs but the majority of them reported having used these models in the past.

We provided participants with a description of the tasks including the goal, the expected outputs and a (fictitious) example of a clinical note. We then asked them to write instructions (in English) for each task with the only constraint being that they had to ensure the model outputs a valid label (for classification tasks) or a list of items (for extraction tasks). Figure 9 (Appendix A.1) shows an example of the instructions given for a classification task.

Initially, we ran a smaller scale pilot study consisting of one classification and one extraction task, and recruited participants who successfully completed the tasks. The process took around 5 hours on average and we compensated each participant at a rate of \$25/hour. We manually reviewed all written instructions and found that some were of poor quality (e.g., did not adhere to the goals of the task, or did not ensure that the model outputs valid responses). In such cases, we removed the author from the study and discarded all of their instructions. We also removed everyone that did not complete all the tasks, resulting in a final collection of instructions from 12 participants. See Appendix A.1 for illustrative examples of the collected instructions<sup>3</sup>.

## 2.3 Models

We measured the performance, variance and fairness of seven general and domain-specific LLMs on each task, using the instructions written by

medical professionals. To assess the impact of clinical instruction tuning, we paired all clinical models with their general domain counterparts. We considered three clinical models: ASCLEPIUS (7B) (Kweon et al., 2023), CLINICAL CAMEL (13B) (Toma et al., 2023), and MEDALPACA (7B) (Han et al., 2023); and their corresponding base models, i.e., LLAMA 2 CHAT (7B), LLAMA 2 CHAT (13B) (Touvron et al., 2023), and ALPACA (7B) (Taori et al., 2023), respectively. We also included MISTRAL IT 0.2 (7B) (Jiang et al., 2023) in our experiments due to its high performance in standard benchmarks.

For all models and datasets, we performed zero-shot inference via prompts with a maximum sequence length of 2048 tokens which included the instruction, the input note, and the output tokens (64 for classification, 256 for extraction). Since most clinical notes were too long to process in a single pass, we followed Huang et al. 2020 and split each note into chunks to be processed independently. For binary classification and prediction tasks, we treated the output for a given input note as positive if at least one of the chunks was predicted to be positive, and negative otherwise. For extraction tasks, we combined the outputs from each chunk into a single set of extractions.

**Evaluation:** Evaluation with generative models is challenging: Models may not respect the desired output format, or may generate responses that are semantically equivalent but lexically different from references (Wadhwa et al., 2023b; Agrawal et al., 2022). We therefore took predictions from the output distribution of the first generated token by selecting the largest magnitude logit from the set of target class tokens. For extraction tasks, we parsed generated outputs and performed exact match comparison with target spans. We report AUROC scores for classification tasks and F1 scores for extraction tasks.

## 3 Results

We present our main results for Mortality Prediction and Drug Extraction in Figure 3 — results for the other classification and information extraction tasks can be found in Appendix A.2, Figures 12 and 13, respectively. Most models show significant variability in performance for alternative but semantically equivalent instructions in both classification and extraction tasks. To further examine these observed disparities, we plotted the distri-

<sup>3</sup>The full set of instructions is available in our code repository



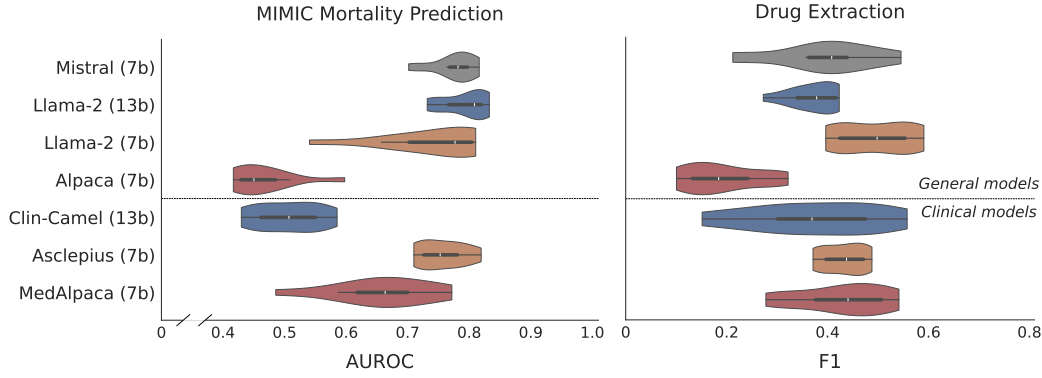


Figure 3: Variability in performance across prompts for the mortality prediction and drug extraction tasks. For most models, different but semantically equivalent prompts yield quite a range of performance.

bution of deltas between the best and worst performing prompts for each task in Figure 2. We see that performance deltas can go up to 0.6 absolute AUROC points for classification tasks and up to 0.4 absolute F1 points for extraction tasks.

In the Mortality Prediction task, we find that LLAMA 2 (13B) outperforms all other models, including the domain-specific ones (Figure 3). However, for the other classification tasks, MISTRAL yields the best results often outperforming the larger models whilst exhibiting less variance (Figure 12). Regarding the clinical models, we observe that ASCLEPIUS consistently attains the best performance in classification tasks albeit with comparable variance.

In the Drug Extraction task, LLAMA 2 (7B) attains the best results on average but with comparable variance to other general LLMs. However, the results for clinical models are mixed: while CLINICAL CAMEL can achieve the highest performance given the best prompt, it also has the highest variance and lowest median performance. MEDALPACA comes close to CLINICAL CAMEL in the best case scenario but with less variance and better median performance. ASCLEPIUS has a median performance similar to that of MEDALPACA but with a much lower variance. We observe similar trends for the other information extraction tasks: LLAMA 2 (7B) consistently outperforms other general LLMs with similar variance, whereas none of the clinical models is clearly superior across tasks — however, ASCLEPIUS seems to have the least variance overall.

To better understand the differences between the general domain and clinical LLMs, we compared their average performance given the best, median and worst prompts. Figures 4 and 5 show the results per model averaged across all classification

and extraction tasks, respectively. Surprisingly, we find that general domain models outperform their domain-specific counterparts — with the exception of ALPACA which performs poorly across all tasks. Again we observe that even though CLINICAL CAMEL *can* outperform its general domain analog in extraction tasks given the best prompt, it also shows more variance and much lower performance in the worst case.

Finally, we investigated whether the observed performance variability can be explained by individual differences between experts in prior experience with LLMs or aptitude in writing effective instructions. To assess this, we measured the performance deltas between each prompt and the median prompt for each classification and extraction task. Figure 6 shows the results for LLAMA 2 (7B) and results for other models can be found in Appendix A.2, figures 14 and 15. We find that there are indeed significant differences at the individual level, both in terms of variance and overall performance, particularly for classification tasks. Only roughly half the users can (somewhat) consistently beat the median performance across tasks. We also note these differences can not be solely explained by prior experience with LLMs — some novice users are able to consistently write more effective instructions as compared to other experienced users. However, one caveat is that this prior experience is most likely with larger commercial models which may be more robust to instruction variations.

### 3.1 Fairness

How do variations in prompt phrasings impact model fairness (here measured as disparities in predictive performance for specific demographic subgroups)? To answer this question, we stratified the patients in the mortality prediction task with

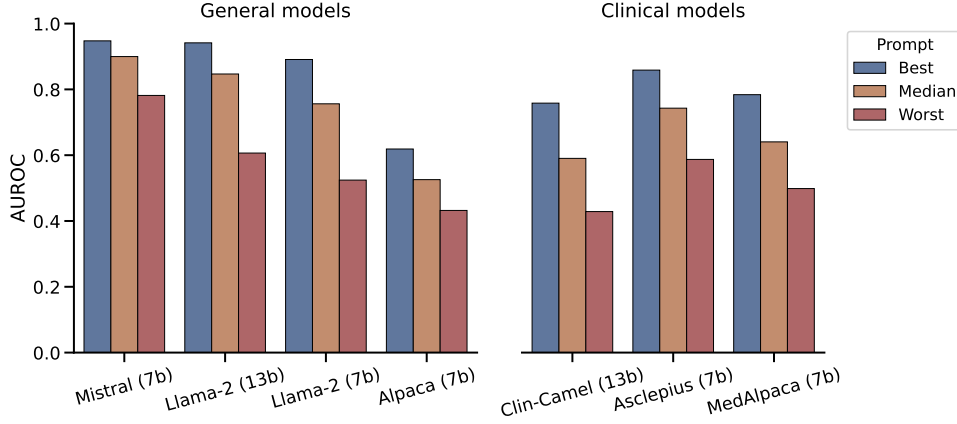


Figure 4: Average AUROC across classification tasks given the best, median, and worst-performing prompts for each model.

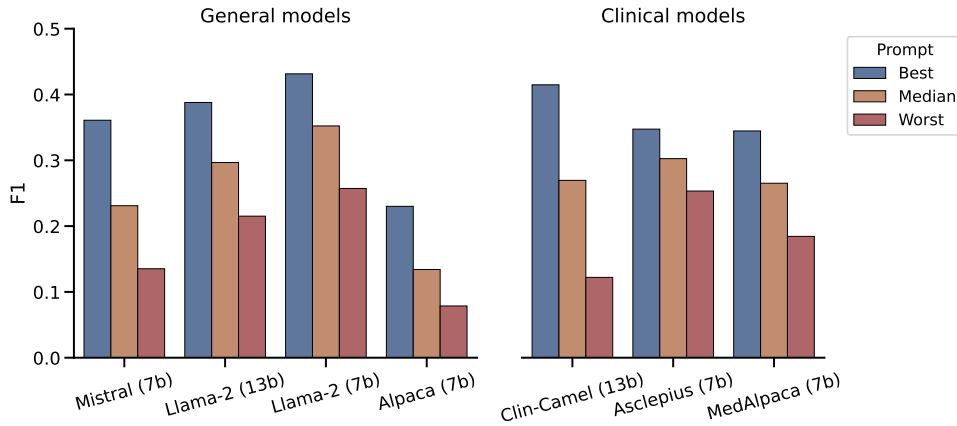


Figure 5: Average F1 across extraction tasks given the best, median, and worst-performing prompts for each model.

		Gender		Total
		Female	Male	
Race	White	52	59	111
	Non-White	24	25	49
Total		76	84	160

Table 2: Distribution of gender and race in the sample used examine model fairness (§3.1)

respect to race and sex. To avoid issues with reliability of performance metrics arising from small sub-samples (Amir et al., 2021) we only consider two broad groups (i.e., *White* and *Non-White*). We sorted the instructions according to their overall performance and plot individual subgroup performance (Figure 7). We repeated the analysis for sex (as indicated in EHR) and present individual subgroup performance in Figure 8.

In line with prior work (Amir et al., 2021; Adam et al., 2022), we observe that models have disparate performance for different subgroups. Both LLAMA 2 (7B) and ASCLEPIUS (7B) tend to under-perform

for non-White patients compared to White counterparts with absolute differences of up to 0.21 and 0.35 AUROC points, respectively. A possible explanation is that the way in which medical staff write clinical notes differ for White vs Black patients (Adam et al., 2022). However, here non-Whites are an heterogeneous group so there may be other confounding factors.

In regards to sex, we again observe noticeable (albeit smaller) differences in performance with LLAMA 2 (7B) performing worse for *Female* patients across all the prompts with relative differences of up to 0.16 absolute AUROC points, and ASCLEPIUS (7B) yielding differences of up to 0.19 points. Overall, these results indicate that natural variations in prompts may translate to wide differences in fairness. Troublingly, a clinician using such models would likely be unaware that apparently benign phrasing changes may disproportionately affect particular demographic groups.

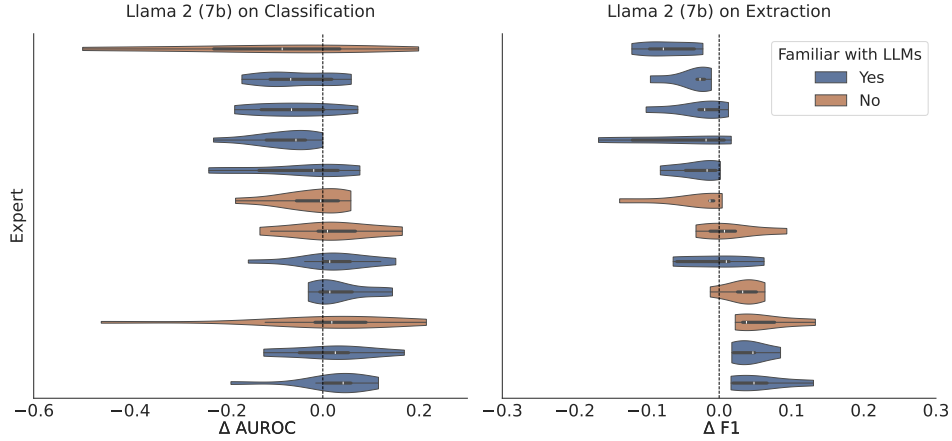


Figure 6: Distribution of performance deltas between each expert’s prompt and the median prompt across all tasks. Each violin plot represents an expert color coded according to their familiarity with LLMs.

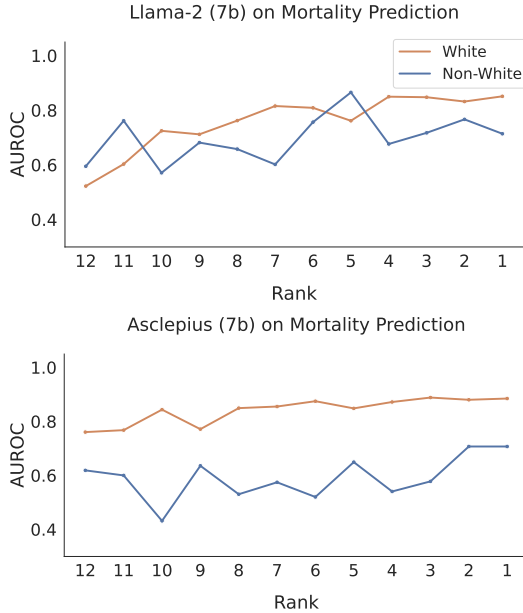


Figure 7: Race subgroup performance on the Mortality Prediction task with a general (top) and clinical model (bottom)

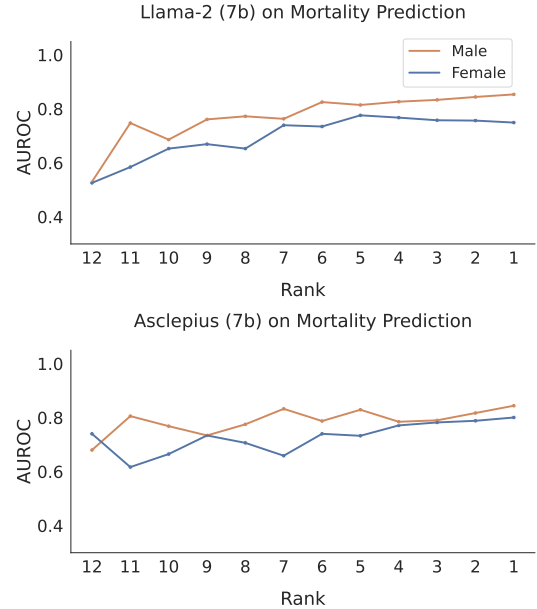


Figure 8: Gender subgroup performance on the Mortality Prediction task with a general (top) and clinical model (bottom)

### 3.2 Discussion

Our experiments show that instruction-tuned LLMs are not robust to plausible variations in instruction phrasings — equivalent but distinct instructions result in significant differences in both task performance and fairness with respect to demographic subgroups. Moreover, we find that no single model yields optimal performance across tasks, e.g. Mistral 7b is the best model for classification but has middling performance in extraction tasks. We also find that general domain models tend to outperform clinical models — although surprising, these findings corroborate prior work on clinical text sum-

marization (Veen et al., 2023). This may be due to the fact that clinical models are fine-tuned with synthetic or proxy data that does not adequately capture the idiosyncrasies of clinical notes from EHR.

## 4 Related Work

**Instruction-following LLMs** Scaling up decoder-only language models imbues them with the ability to solve various tasks given only instructions or a small set of examples at inference time (Brown et al., 2020; Chowdhery et al., 2022). Follow-up work sought to improve this by explicitly training GPT-3 to follow instructions

and provide helpful and harmless responses via Reinforcement Learning from Human Feedback (Ouyang et al., 2022; OpenAI, 2022). Others showed that fine-tuning with a causal language modeling objective over labeled data formatted as instruction/response pairs is sufficient to endow even (comparatively) smaller models with instruction-following abilities (Sanh et al., 2021; Wei et al., 2021). This motivated extensive work on compiling large instruction-tuning datasets, such as the Flan 2021 (Chung et al., 2022) and Super-NaturalInstructions collections (Wang et al., 2022), each encompassing over 1600 NLP tasks, and OPT-IML collection with 2000 tasks (Iyer et al., 2022).

**LLM Prompt Sensitivity** However, LLMs are sensitive to how prompts are constructed (Tjua et al., 2023; Raj et al., 2023). In few-shot learning, factors such as the prompt format (Sclar et al., 2023; Chakraborty et al., 2023), as well as the choice (Gutiérrez et al., 2022) and ordering (Lu et al., 2022; Pezeshkpour and Hruschka, 2023) of exemplars have a significant impact on task performance. In zero-shot settings, Webson and Pavlick (2022) found that models often realize similar performance with misleading or irrelevant prompts as with correct ones. Elsewhere, Sun et al. (2023) showed that general domain instruction-tuned LLMs are not robust to variations in instructions — specifically, they found that models underperform when given novel instructions unseen in training. Our work contributes to this line of research by focusing on the clinical domain.

**LLMs for Clinical Tasks** General domain LLMs encode a surprising amount of clinical and biomedical knowledge allowing them to solve various prediction and information extraction tasks via natural language instructions (Singhal et al., 2023; Agrawal et al., 2022; Munnangi et al., 2024). However, smaller models fine-tuned on task-specific data can outperform generalist LLMs in clinical tasks (Lehman et al., 2023). At the same time, there is a dearth of large high-quality clinical text datasets to train LLMs due to privacy considerations. Researchers have tried to overcome this by exploiting synthetic data generated from biomedical and clinical literature and question answering datasets to train domain-specific models (Toma et al., 2023; Kweon et al., 2023; Han et al., 2023). However, the resulting models are often outperformed by general domain variants (Veen et al.,

2023; Excoffier et al., 2024) — our experimental results confirm these observations.

In a contemporaneous study Chang et al. (2024) convened a panel of 80 multidisciplinary experts to red team ChatGPT models for the appropriateness of the responses in medical use cases. Experts were asked to write (non-adversarial) prompts for clinically relevant scenarios and the responses were judged by medical doctors with respect to safety, privacy, hallucinations, and bias. This work is complementary to ours in that it aims to stress test models for the *appropriateness* of their responses to healthcare related prompts whereas we focus on their *sensitivity* to prompt variations.

## 5 Conclusions

This paper presents a large-scale evaluation of instruction-tuned open-source LLMs for clinical classification and information extraction tasks on clinical notes (from EHR). We specifically focus on model robustness to natural differences in prompts written by medical professionals. We recruited 12 practitioners with different professional and demographic backgrounds, medical specialties, and years of experience to write prompts for 16 clinical tasks spanning binary classification, outcome prediction, and information extraction.

There are a few main generalizable takeaways relevant to machine learning in healthcare in this work. First, the performance LLMs realize on the same clinical task varies substantially across prompts written by different domain experts, and this holds across all models. Second, the domain-specific (clinical) models we evaluated perform, in general, worse than their general domain counterparts. Third, prompt variations have concerning implications for fairness — we find that alternative prompts yield different levels of fairness. Based on these findings, we recommend that practitioners exercise caution when using instruction-tuned LLMs for high stakes clinical tasks which may ultimately impact patient health. Crucially, clinicians using LLMs should be made aware that subtle, plausible variations in phrasings may yield quite different outputs. Beyond healthcare, this work enriches our understanding of (the lack of) LLM robustness and—we hope—will motivate research into new methods to improve models in this respect.



## 6 Limitations

Our study reveals that open-source instruction-tuned LLMs are sensitive to instruction phrasings and suggests caution in adopting these models for applications that may impact personal health and well-being. However, this work has several limitations. First, we acknowledge that our findings may not generalize to larger commercial models but cost and privacy considerations may preclude the deployment of proprietary models for real-world healthcare applications. Second, we endeavored to recruit a diverse group of medical professionals but our final pool of participants may not be a representative sample of the potential users of these technologies. Moreover, participants were not allowed to see the results of their instructions but in the real world users would have the opportunity to experiment with different prompts and learn how to best use these models. Third, our evaluation protocol for classification tasks may not reflect real world usage — we induced model predictions from the logit distribution of the first generated token. However, in practice users can only see the final generated outputs and must be able to parse and interpret these in the context of the task at hand. Finally, our analysis showed that variations in instructions have implications for fairness with respect to race and gender. However, we did not examine the impact of these disparities on intersectional identities which are often affected by compounded biases.

## Acknowledgments

This work was supported in part by National Science Foundation (NSF) award 1901117, and by the National Institutes of Health (NIH) award R01LM013772.

We also thank the reviewers, for their valuable feedback and comments that helped improve this work.

## References

Hammad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. [Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations](#). In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22*. ACM.

Monica Agrawal, Stefan Hegselmann, Hunter Lang,

Yoon Kim, and David Sontag. 2022. [Large language models are few-shot clinical information extractors](#). *Preprint*, arXiv:2205.12689.

Silvio Amir, Jan-Willem van de Meent, and Byron C. Wallace. 2021. [On the impact of random seeds on the fairness of clinical classifiers](#). *Preprint*, arXiv:2104.06338.

Masaki Asada and Ken Fukuda. 2024. Enhancing relation extraction from biomedical texts by large language models. In *International Conference on Human-Computer Interaction*, pages 3–14. Springer.

Dhananjay Ashok and Zachary C. Lipton. 2023. [Promptner: Prompting for named entity recognition](#). *Preprint*, arXiv:2305.15444.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. [Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5698–5711, Toronto, Canada. Association for Computational Linguistics.

Crystal T. Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A. Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H. Jacobson, Binglan Li, Celeste H. Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A. Fries, Jiapeng Xu, Joe Jamison, John S. Tamaresis, Jonathan H Chen, Joshua Lazaro, Juan M. Banda, Julie J. Lee, Karen Ebert Matthys, Kirsten R. Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T. Truong, Scott L. Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y. Eulalio, Wenyan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam H. Shah, and Roxana Daneshjou. 2024. [Red teaming large language models in medicine: Real-world insights on model behavior](#). *medRxiv*.

Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. [Can AI Help Reduce Disparities in General](#)

- [Medical and Mental Health Care?](#) *AMA journal of ethics*, 21 2:E167–179.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Jean-Baptiste Excoffier, Tom Roehr, Alexei Figuerola, Jens-Michalis Papaioannou, Keno Bressem, and Matthieu Ortala. 2024. [Generalist embedding models are better at short-context clinical semantic search than specialized embedding models](#). *Preprint*, arxiv:2401.01943.
- Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. MedAlpaca—An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247*.
- Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and A. G. Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6.
- Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. [2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records](#). *Journal of the American Medical Informatics Association: JAMIA*, 27(1):3–12.
- Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. [Clinicalbert: Modeling clinical notes and predicting hospital readmission](#). *Preprint*, arXiv:1904.05342.
- Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Guochao Jiang, Ziqin Luo, Yuchen Shi, Dixuan Wang, Jiaqing Liang, and Deqing Yang. 2024. Toner: Type-oriented named entity recognition with generative language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16251–16262.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.
- Sunjun Kweon, Junu Kim, Jiyoung Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. 2023. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*.
- Joon Lee, Daniel J. Scott, Mauricio Villarroel, Gari D. Clifford, Mohammed Saeed, and Roger G. Mark. 2011. [Open-access mimic-ii database for intensive care research](#). In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8315–8318.
- Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Szolovits Peter Alsentzer, Emily, Alistair Johnson, Emily Alsentzer, et al. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Monica Munnangi, Sergey Feldman, Byron C Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. [On-the-fly definition augmentation of llms for biomedical ner](#). *Preprint*, arXiv:2404.00152.
- Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.
- OpenAI. 2022. [ChatGPT-3.5](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Ridam Pal, Hardik Garg, Shashwat Patel, and Tavpritesh Sethi. 2023. Bias amplification in intersectional subpopulations for clinical phenotyping by large language models. *medRxiv*, pages 2023–03.
- Jon Patrick and Min Li. 2010. [High Accuracy Information Extraction of Medication Information from Clinical Notes: 2009 I2b2 Medication Extraction Challenge](#). *Journal of the American Medical Informatics Association : JAMIA*, 17(5):524–527.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. [Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions](#). *arXiv preprint*. ArXiv:2308.11483 [cs].
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. [Semantic Consistency for Assuring Reliability of Large Language Models](#). *arXiv preprint*. ArXiv:2308.09138 [cs].
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. [Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting](#). *arXiv preprint*. ArXiv:2310.11324 [cs].
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Amber Stubbs, Christopher Kotfila, Hua Xu, and Ozlem Uzuner. 2015. [Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2](#). *Journal of biomedical informatics*, 58(Suppl):S67.
- Amber Stubbs and Özlem Uzuner. 2019. [New approaches to cohort selection](#). *Journal of the American Medical Informatics Association*, 26(11):1161–1162.
- Jiuding Sun, Chantal Shaib, and Byron C. Wallace. 2023. [Evaluating the Zero-shot Robustness of Instruction-tuned Language Models](#). *arXiv preprint*. ArXiv:2306.11270 [cs].
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Lindia Tjuaatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. [Do LLMs exhibit human-like response biases? A case study in survey design](#). *arXiv preprint*. ArXiv:2311.04076 [cs].
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint arXiv:2305.12031*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Özlem Uzuner. 2009. [Recognizing obesity and comorbidities in sparse data](#). *Journal of the American Medical Informatics Association*, 16(4):561–570.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. [2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text](#). *Journal of the American Medical Informatics Association*, 18(5):552–556.
- Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Blüthgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curt P. Langlotz, Jason Hom, Sergios Gatidis, John M. Pauly, and Akshay S. Chaudhari. 2023. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature Medicine*, 30:1134–1142.
- Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023a. [Revisiting relation extraction in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.
- Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C Wallace. 2023b. Jointly extracting interventions, outcomes, and findings from RCT reports with LLMs. In *Machine Learning for Healthcare Conference*, pages 754–771. PMLR.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Albert Webson and Ellie Pavlick. 2022. [Do Prompt-Based Models Really Understand the Meaning of Their Prompts?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

## A Appendix

### A.1 Instruction Collection

To collect instructions from experts, we provided them with a description of the tasks including the goal, the expected outputs and a (fictitious) example of a clinical note. Figure 9 is an example of the instructions given for a classification task; and Figures 10 and 11 show examples of collected instructions. We released the full set of collected instructions along with code.

### A.2 Results

In this section we present additional results from our experiments. We show detailed results in terms of the mean performance and standard deviation for all the classification and information extraction tasks in tables 3 and 4, respectively.

Figures 12 and 13 plot the variability in performance across classification and extraction tasks, respectively. Figures 14 and 15 plot the deltas in performance between individual expert’s prompts and the median prompt per task, for general domain and clinical models, respectively.

Figure 16 show race subgroup performance for the Mortality Prediction task for all the models, and Figure 17 shows a similar analysis for sex.

Our overall results show that, in general, different prompt phrasings yield different performance. Are there prompts that are consistently effective across models? To investigate this, we ranked each prompt with respect to the performance and calculated the median across models. Figures 18 and 19 depict the median performance ranking (among all 12 prompts) achieved by the instructions written by each expert. For classification tasks such as Cohort Abdominal and Cohort Make Decisions, Expert 7 wrote prompts that are consistently among the best performing ones for most models, which is also the case for the prompts written by Expert 11 across five classification tasks. On the other hand, prompts from Expert 2 were consistently among the lower performing ones. A similar pattern can be seen in the extraction tasks, where Experts 6 and 8 wrote some of the best-performing prompts for most of these tasks. This suggests that, to an extent, the performance of prompts is consistent even when tested on different models.



**Your Job**

Write down an instruction you would use to prompt the LLMs to do the following task.

**Task Description**

Given some notes from a patient encounter, classify whether or not the patient will die during the course of a stay in the hospital given notes from the first 48 hours.

```

...
ADMISSION DATE:
5-29-94
DISCHARGE DATE:
7-3-94
HISTORY OF PRESENT ILLNESS:
The patient is a 58-year-old right-hand dominant white female with a long history of hypertension; changed her
medications from Clonidine to Aldomet six weeks ago...
MEDICATIONS ON ADMISSION:
Vasotec 40 mg q.day, Soma 1 tablet q.day, Demerolprn, Clonidine.
SOCIAL HISTORY:
The patient lives with two people in Cinglenda.
PHYSICAL EXAMINATION:
On physical examination, the patient is in no acute distress, afebrile, blood pressure 135/82, heart rate 82 and
regular, and no bruits ...
TR :
[ report_end ]

```

**Your Prompt**

Figure 9: Example of instructions for annotators for a classification task; we provided participants with a description of the tasks including the goal, the expected outputs and a (fictitious) example of a clinical note.

**TASK: Cohort Drug Abuse**

**CLINICAL NOTE:** <note>

**INSTRUCTION:**

Based on the medical records provided, answer the question. Does this patient meet the definition of drug abuse? Respond "Yes" if the patient meets the definition of drug abuse, or respond "No" if the patient does not meet the definition of drug abuse.

Figure 10: Example of instructions for 'Cohort drug abuse' classification task.

<p><b>TASK: Mortality Prediction</b></p> <p><b>CLINICAL NOTE:</b> &lt;note&gt;</p> <p><b>INSTRUCTION:</b></p> <p>Based on the medical notes provided from the first 48 hours of the patient\'s hospital stay, please classify whether the patient will die during their hospital stay. Respond with either "Yes" if the patient will die during their stay in the hospital or "No" if the patient will not die during their stay in the hospital.</p>
---

Figure 11: Example of instructions for ‘Mortality Prediction’ classification task.

<b>Model / Dataset</b>	<b>MISTRAL IT 0.2 (7B)</b>	<b>LLAMA 2 CHAT (13B)</b>	<b>LLAMA 2 CHAT (7B)</b>	<b>ALPACA (7B)</b>	<b>CLINICAL CAMEL (13B)</b>	<b>ASCLEPIUS (7B)</b>	<b>MEDALPACA (7B)</b>
Obesity Co-Morbidity (Asthma)	0.974 ±(0.014)	0.908 ±(0.111)	0.696 ±(0.145)	0.479 ±(0.017)	0.594 ±(0.059)	0.732 ±(0.086)	0.557 ±(0.078)
Cohort Alcohol Abuse	0.980 ±(0.028)	0.898 ±(0.142)	0.836 ±(0.148)	0.549 ±(0.126)	0.517 ±(0.177)	0.894 ±(0.084)	0.715 ±(0.146)
Obesity Co-Morbidity CAD	0.963 ±(0.017)	0.933 ±(0.067)	0.796 ±(0.096)	0.512 ±(0.033)	0.649 ±(0.107)	0.702 ±(0.154)	0.679 ±(0.071)
Cohort Drug Abuse	0.941 ±(0.039)	0.923 ±(0.04)	0.934 ±(0.048)	0.570 ±(0.132)	0.698 ±(0.138)	0.938 ±(0.042)	0.756 ±(0.119)
Cohort English	0.974 ±(0.055)	0.824 ±(0.123)	0.790 ±(0.165)	0.460 ±(0.071)	0.586 ±(0.076)	0.737 ±(0.078)	0.552 ±(0.058)
Cohort Make Decision	0.709 ±(0.178)	0.623 ±(0.238)	0.710 ±(0.171)	0.644 ±(0.047)	0.597 ±(0.174)	0.817 ±(0.074)	0.513 ±(0.098)
Cohort Abdominal	0.750 ±(0.034)	0.707 ±(0.076)	0.644 ±(0.034)	0.483 ±(0.029)	0.506 ±(0.069)	0.637 ±(0.052)	0.648 ±(0.059)
Obesity Co-Morbidity (Diabetes)	0.987 ±(0.011)	0.958 ±(0.063)	0.775 ±(0.114)	0.560 ±(0.041)	0.637 ±(0.109)	0.762 ±(0.124)	0.686 ±(0.05)
Obesity Classification	0.943 ±(0.05)	0.9 ±(0.087)	0.639 ±(0.113)	0.534 ±(0.03)	0.612 ±(0.074)	0.453 ±(0.177)	0.64 ±(0.084)
Mortality Prediction	0.777 ±(0.034)	0.794 ±(0.036)	0.742 ±(0.083)	0.466 ±(0.051)	0.506 ±(0.052)	0.757 ±(0.037)	0.658 ±(0.08)

Table 3: Mean and Standard Deviation for instructions on classification tasks across all models and all tasks

<b>Model / Dataset</b>	<b>MISTRAL IT 0.2 (7B)</b>	<b>LLAMA 2 CHAT (13B)</b>	<b>LLAMA 2 CHAT (7B)</b>	<b>ALPACA (7B)</b>	<b>CLINICAL CAMEL (13B)</b>	<b>ASCLEPIUS (7B)</b>	<b>MEDALPACA (7B)</b>
Medication Extraction	0.351 $\pm(0.111)$	0.559 $\pm(0.072)$	0.608 $\pm(0.084)$	0.231 $\pm(0.069)$	0.509 $\pm(0.15)$	0.562 $\pm(0.027)$	0.529 $\pm(0.047)$
Concept Problem Extraction	0.265 $\pm(0.051)$	0.325 $\pm(0.035)$	0.329 $\pm(0.027)$	0.131 $\pm(0.029)$	0.3 $\pm(0.035)$	0.256 $\pm(0.019)$	0.229 $\pm(0.021)$
Concept Test Extraction	0.154 $\pm(0.076)$	0.197 $\pm(0.066)$	0.236 $\pm(0.05)$	0.097 $\pm(0.025)$	0.117 $\pm(0.078)$	0.194 $\pm(0.025)$	0.109 $\pm(0.049)$
Concept Treatment Extraction	0.165 $\pm(0.084)$	0.244 $\pm(0.086)$	0.367 $\pm(0.093)$	0.086 $\pm(0.031)$	0.198 $\pm(0.129)$	0.308 $\pm(0.039)$	0.193 $\pm(0.072)$
Drug Extraction	0.394 $\pm(0.101)$	0.373 $\pm(0.047)$	0.495 $\pm(0.072)$	0.192 $\pm(0.074)$	0.372 $\pm(0.128)$	0.432 $\pm(0.042)$	0.429 $\pm(0.086)$
Risk Factor CAD Extraction	0.057 $\pm(0.009)$	0.081 $\pm(0.018)$	0.079 $\pm(0.024)$	0.067 $\pm(0.056)$	0.122 $\pm(0.046)$	0.063 $\pm(0.012)$	0.103 $\pm(0.029)$

Table 4: Mean and Standard Deviation for instructions on extraction tasks across all models and all tasks

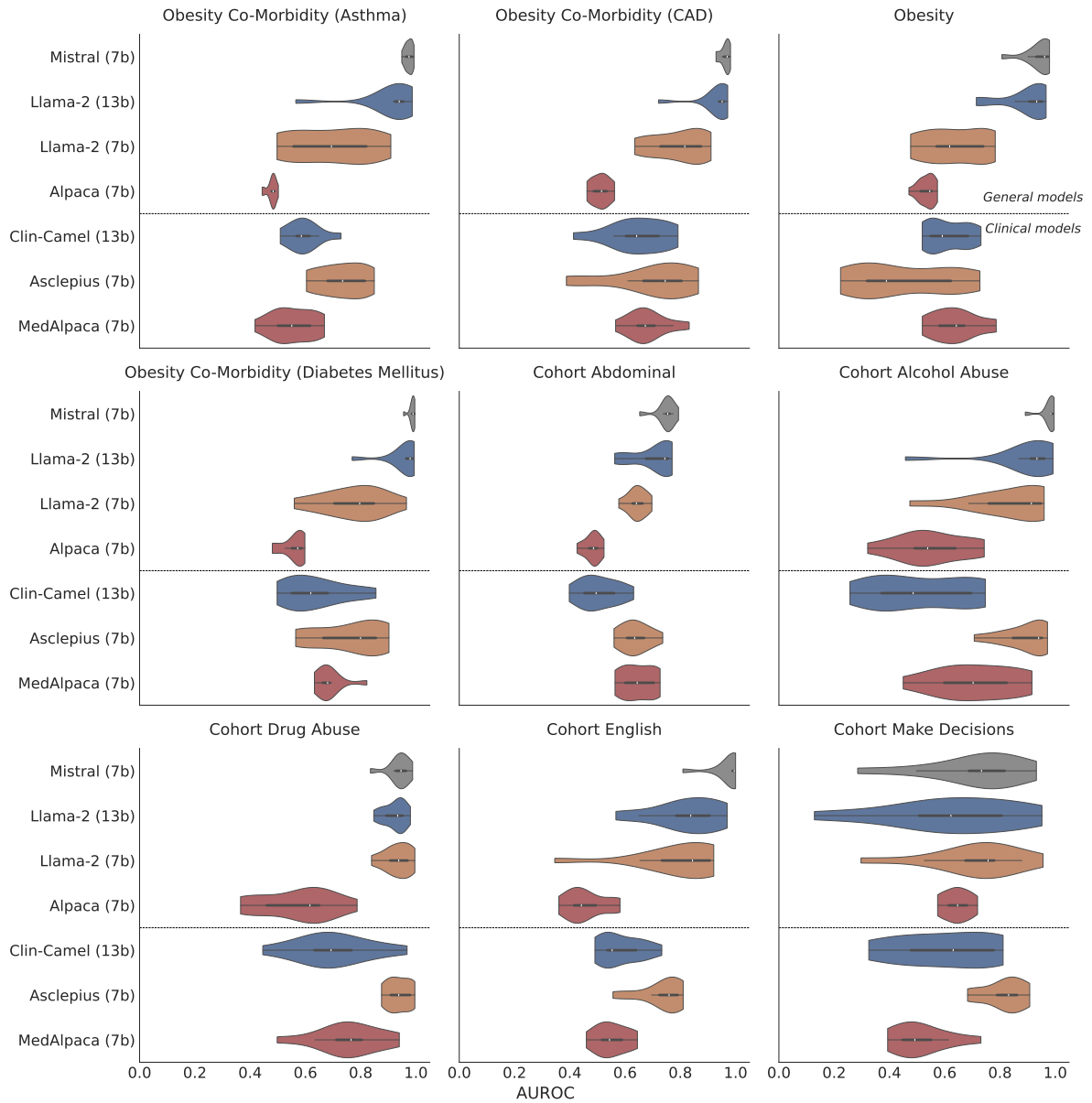


Figure 12: Variability in performance across prompts for binary classification tasks. Again we observe that different (equivalent) instructions yield wide variances in performance, suggesting an undue sensitivity to phrasings.

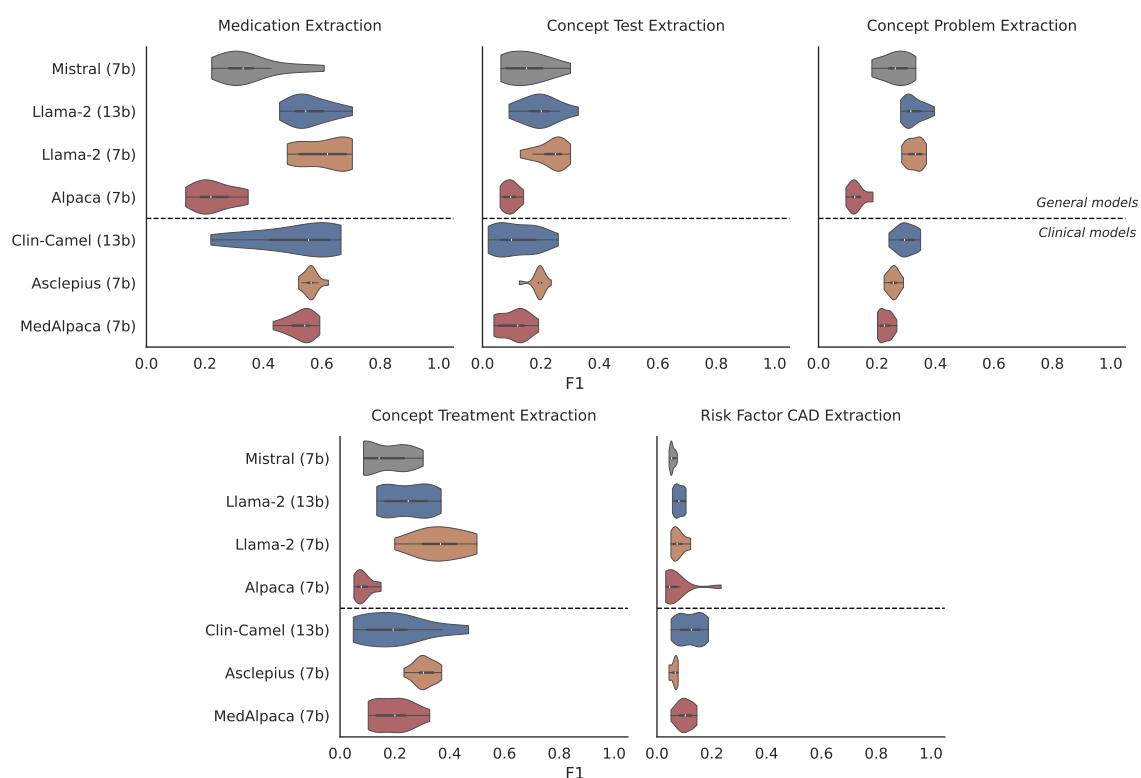


Figure 13: Variability in performance across prompts for the remaining 5 extraction tasks. As mentioned, for most models, different but semantically equivalent prompts yield quite a range of performance.

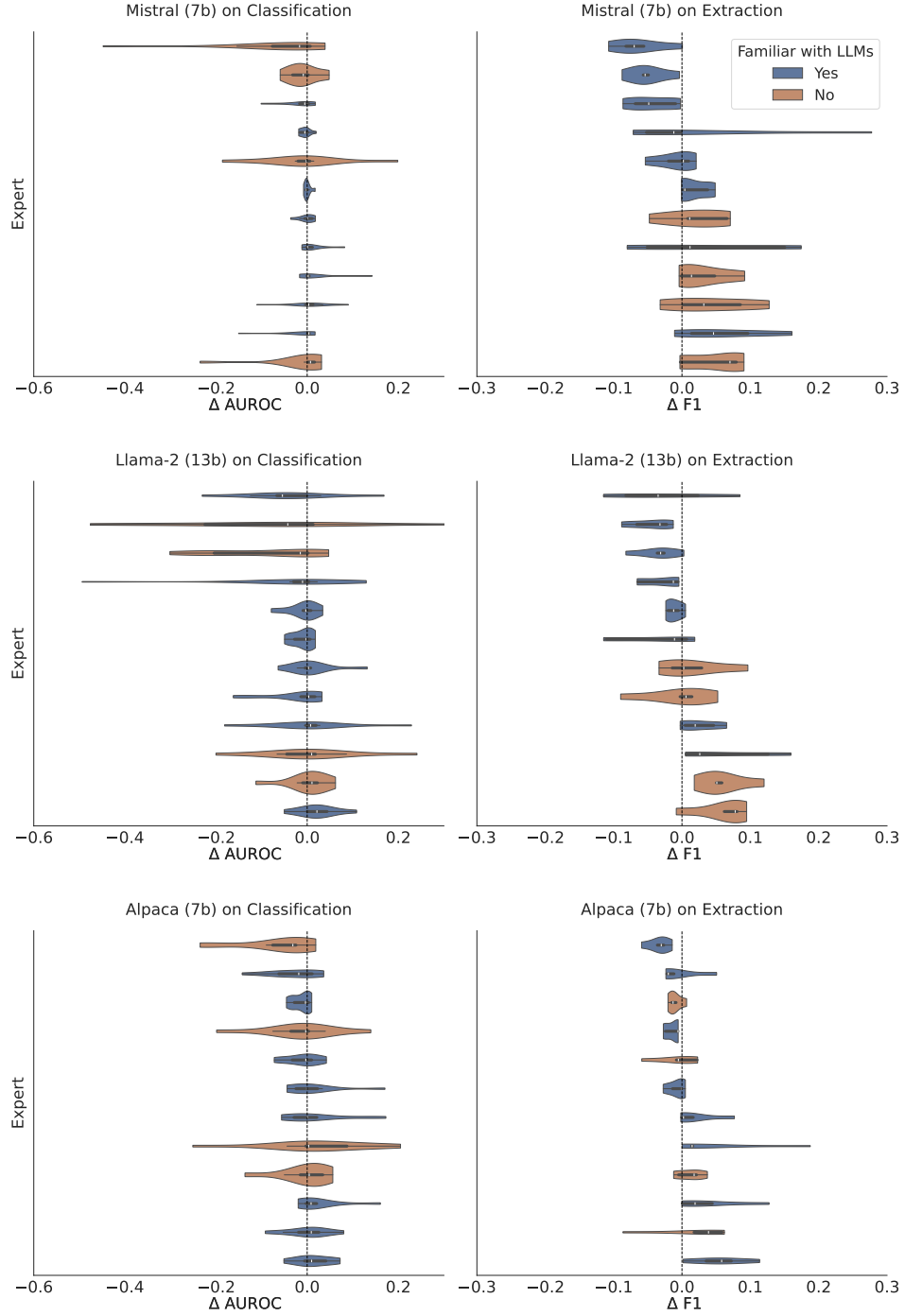


Figure 14: Distribution of performance deltas between each expert’s prompt and the median prompt across all tasks for each *general* model. Each violin plot represents an expert color-coded according to their familiarity with LLM.



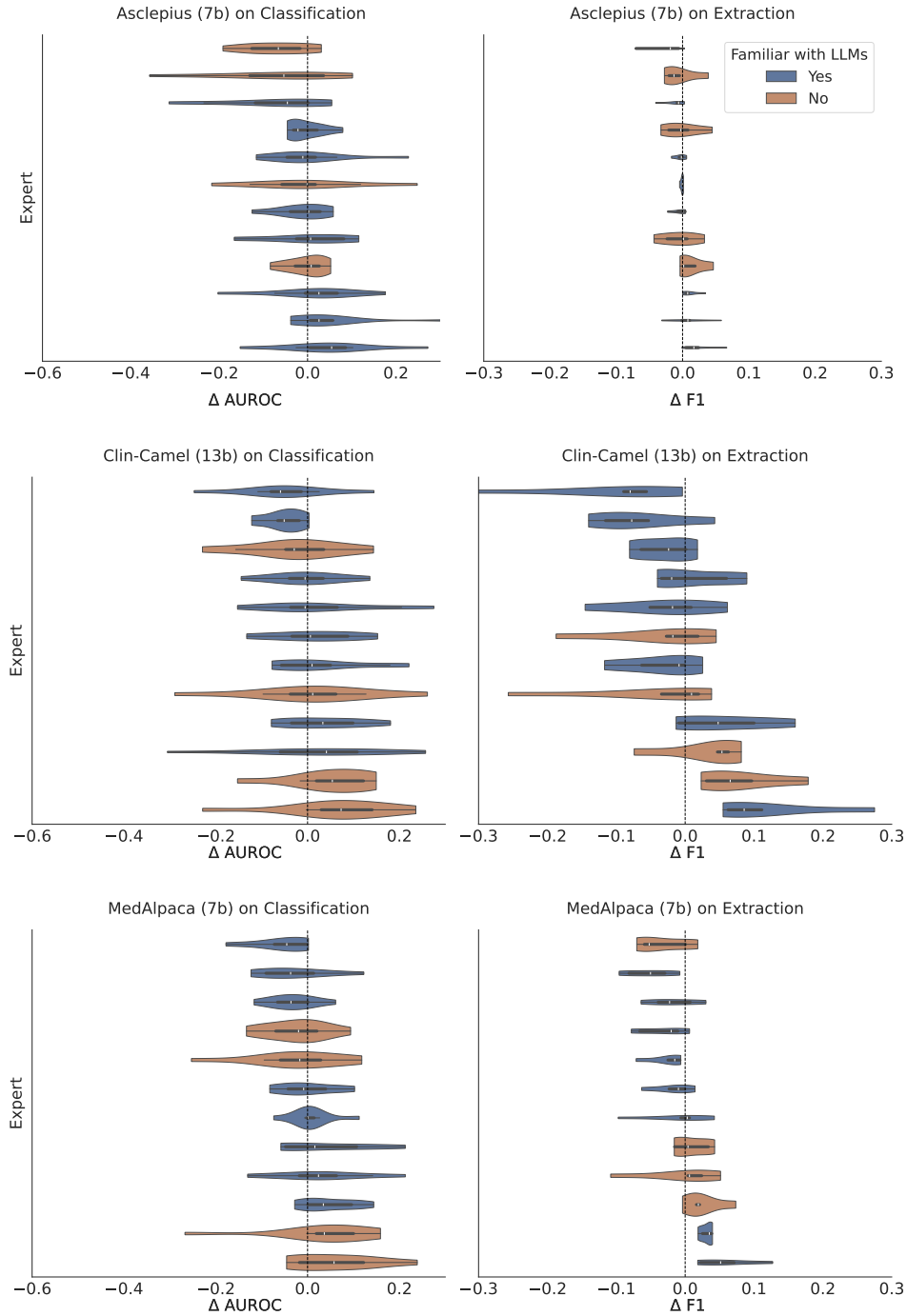


Figure 15: Distribution of performance deltas between each expert’s prompt and the median prompt across all tasks for each *clinical* model. Each violin plot represents an expert color-coded according to their familiarity with LLM.

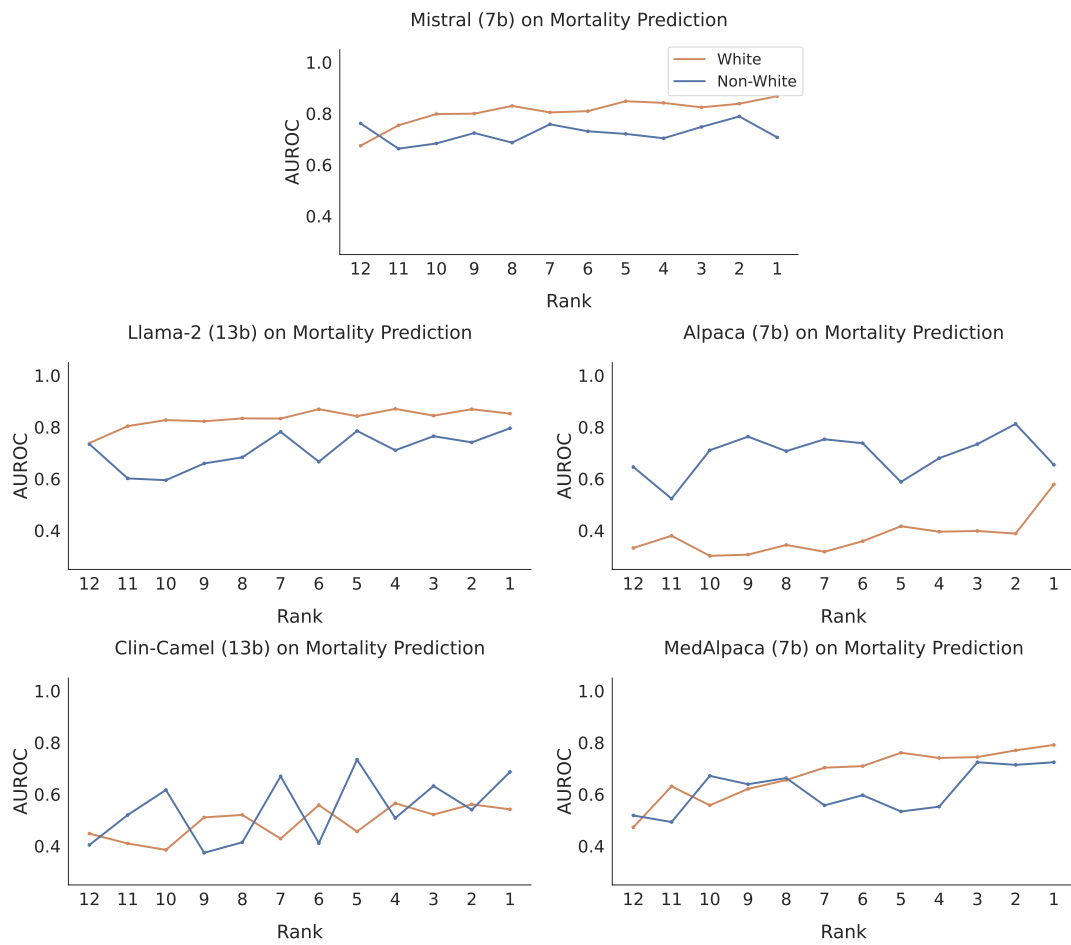


Figure 16: Race subgroup performance on the Mortality Prediction task with a general (left) and clinical model (right). Mistral has no clinical counterpart in our study.

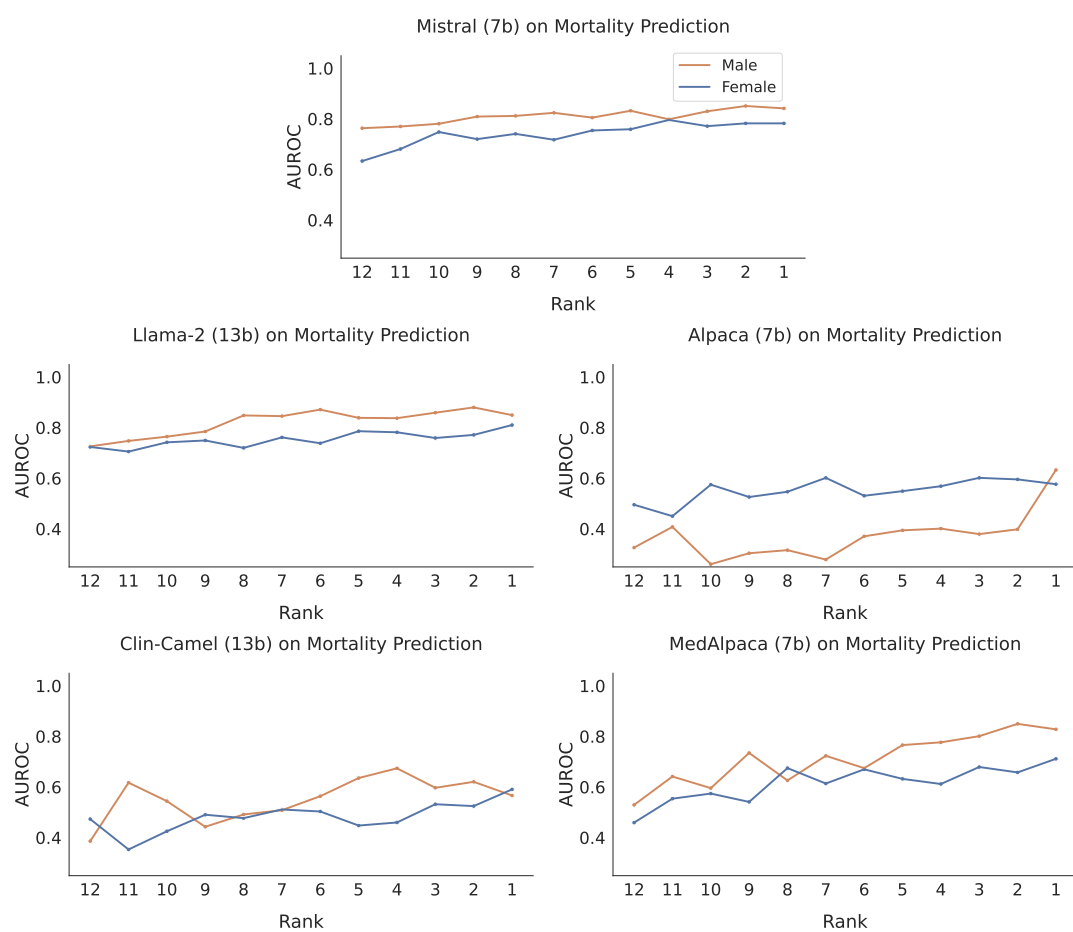


Figure 17: Sex subgroup performance on the Mortality Prediction task with a general (left) and clinical model (right). Mistral has no clinical counterpart in our study.

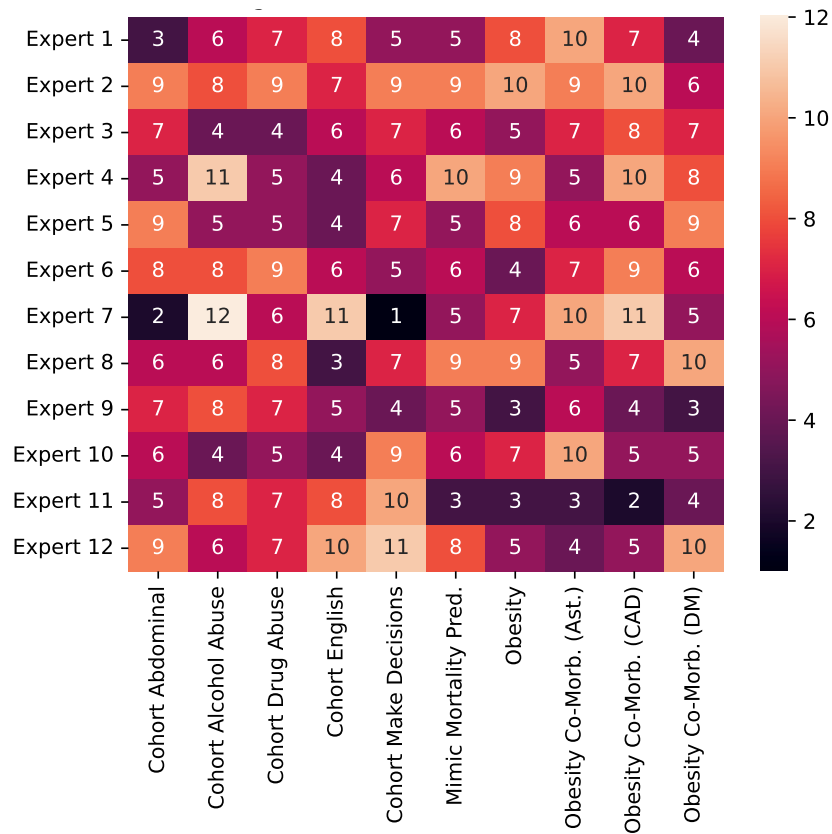


Figure 18: Median ranking of prompts written by experts for classification tasks across models.

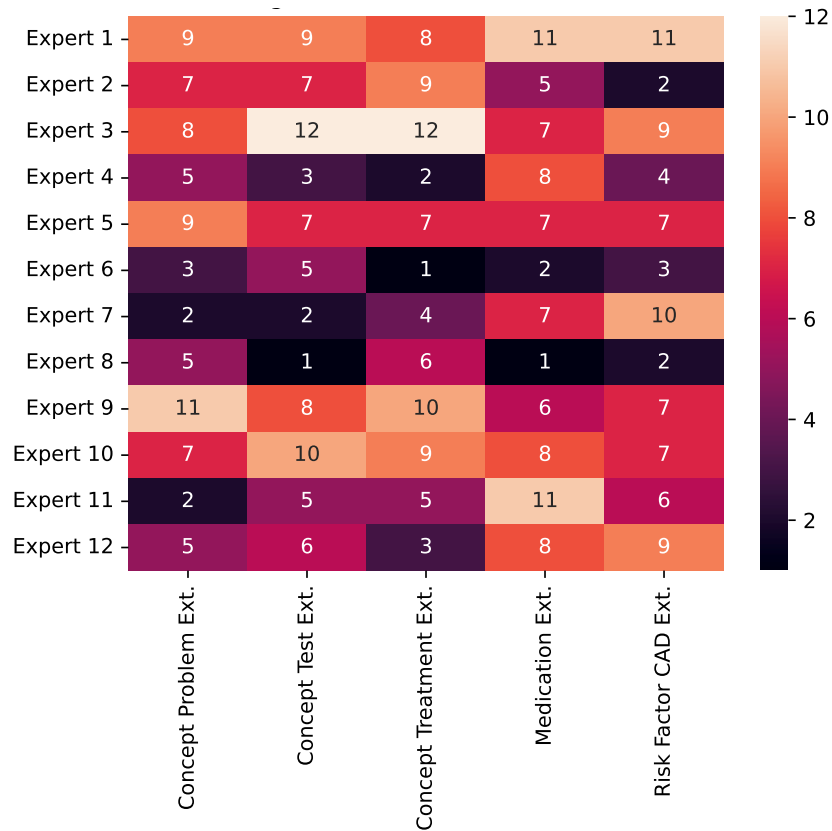


Figure 19: Median ranking of prompts written by experts for extraction tasks across models.